MORNINGSTAR®

# Morningstar Equity Comparables Methodology

**Morningstar Quantitative Research**
17 October 2018
Version 1.0

**Authors**
Patrick Caldon
Director of Quantitative Research
+61 2 9276 4473
patrick.caldon@morningstar.com

Taylor Hess
Senior Quantitative Analyst
+1 312 696 6165
taylor.hess@morningstar.com

## Executive Summary

Morningstar developed the Morningstar Equity Comparables system to give investors and financial professionals an objective benchmark for comparing companies. Morningstar Equity Comparables is genuinely different to other industry classification schemes. We start from the bottom up with comparable companies, as opposed to the top down with sector definitions. For every pair of companies, we determine how similar they are–anywhere from closely comparable to distantly related based on automated analysis of the companies' own business description. We automatically analyse the text of the business description and work out whether companies are talking about similar things as they describe their businesses. Businesses described in similar terms are comparable.

### Key Takeaways

- Morningstar Equity Comparables is a bottom up, text-based company similarity scheme
- This model provides flexible, cross-sector similarity scores

## Thinking Outside the Hierarchy

How should we classify businesses? How do you justify every decision in constructing a taxonomy? How do you avoid catch-all "other" categories? How do you know the classification is correct in some sense? Our argument is that any taxonomy of business will end up telling us more about the person doing the classification than the objects being classified. We believe the problems with hierarchical classification lead to poorer results when it is employed in valuations or risk management.

In this document we present a better technique, which we call Morningstar Equity Comparables, based on a graph or network of business similarity. We will show how to automatically analyze the text of filings and company disclosures and automatically calculate the similarity of two firms from this document-based fundamental data. We do not employ data analysts to examine businesses with the subjectivity that entails.

It is natural for people to categorize objects in hierarchies or taxonomies–a top-down classification that operates by a mechanism of class inclusion. The General Industry Classification System, or GICS, from MSCI and Standard and Poor's is perhaps the best-known such system today, but many similar systems

exist, and they are pervasive in the finance industry. We will proceed to criticize GICS because we expect it to be most familiar to the reader, but our criticism applies equally to all top-down taxonomic business classifications, including SIC, NAICS, and Morningstar's own Global Equity Classification Standard, or GECS. Presently in GICS, the universe of businesses is divided into 10 sectors, 24 industry groups, 68 industries, and 154 subindustries. Once an industry is placed in a subindustry–for example, Peabody Coal common stock might be placed in the coal subindustry–because of the strict hierarchy in the classification, it must also be placed in the successive higher levels of the hierarchy, such as the oil, gas, and consumable fuels industry, the energy industry group, and the energy sector.

But why look at the similarity of businesses anyway? Similar businesses will tend to have similar cash flows, which market prices implicitly discount at a similar rate, which gives a measurement of the comparability of businesses two broad purposes. First, valuation: Given a business, an analyst may wish to understand similar businesses for a comparable business-based relative valuation. Second, risk management: Uncertain future cash flows from similar businesses will tend to be correlated, which will make the market returns of their securities correlated, and many portfolio construction methodologies discourage investing in assets with correlated returns.

Our system can directly capture the similarity of firms in what would be different sectors in hierarchical systems. While we criticize hierarchical systems for their poor performance, they still have the advantage of being simple to understand, and once the hierarchy is defined it becomes straightforward to determine where an industry is in terms of other industries. By constructing a fragment of a "universal language"[1] to describe businesses, an investor can at least simplify the problem of choice and process the vast amounts of information efficiently[2]. For instance, with GICS we can infer very simply that coal companies are energy companies, closely related to other oil and gas companies. For indexing purposes–or building an exchange-traded fund–this means that an "Energy Index" has a clear definition. We think hierarchical systems are superior to ours in this indexing and ETF role.

But these bright boundaries between groups and subgroups are any hierarchical system's Achilles heel. The difficulty with this particular example of coal–and we argue for many examples–is that these categorical judgements become a problem around the boundaries. The world of business is not so simple as to be easily squeezed into a hierarchy. For instance, there are two uses of coal: coking coal and thermal coal. Coking coal is used for steel making (materials in GICS), thermal coal for electricity and heating (energy in GICS). Furthermore, coal is mined, which suggests the same cost structure as a mining company. So, the classification of Peabody Coal as an energy stock is a bit arbitrary. This hierarchy suggests that it's equally wise to diversify the risk in coal with steel businesses as with healthcare businesses: It can easily be seen how naïve application of the classification system might lead to poor investment choices. Every hierarchical classification system with clear-cut boundaries will be weak around these kinds of corner cases. We don't suggest that GICS is wrong to put coal in the energy sector. We argue instead that employing a hierarchical taxonomy leads inevitably to boundary

1 The Analytical Language of John Wilkins, Jorge Luis Borges.
2 http://faculty.som.yale.edu/nicholasbarberis/jfe_final.pdf

cases where the industry classification becomes arbitrary, so for a material portion of the subclasses or businesses there is no clear final place where the subclass or a business should sit within the hierarchy.

## Morningstar Equity Comparables Overview

Morningstar Equity Comparables is an automated business similarity measurement scheme based on an analysis of the description of the business by the business' own management, as well as other text used to describe the company in question. The assumption is that the management of two businesses that do roughly the same thing will describe their businesses in the same terms.

Morningstar Equity Comparables covers approximately 14,000 firms, limited by the amount of English text we can get that talks about these companies. In practice, this means our coverage substantially consists of companies that trade in US markets and largely companies with higher market capitalizations. This increases and falls somewhat as firms list and delist. The system is based on the most recent annual report, Wikipedia text data, and analyst reports. We regenerate our similarity score every other week and include all relevant documents since the prior run.

For each calculation, we start by automatically reading the collection of recent company disclosure documents–10-Ks, annual reports, prospectuses–and associating them with the company itself. This is also the point at which we grab Wikipedia and analyst report data. The system automatically finds the parts of the documents that discuss the business. In this manner, we build a representation of the documents that are convenient for further processing.

We then use a formal model of how management goes about writing a disclosure document. This is based on the notion of a topic model. A topic model is a generative model for an agent writing a document. In the topic model we employ, an agent–in our case company management writing disclosure documents–is modelled as an automaton randomly selecting words from a collection of word groups, called topics. Each topic is a group of words that might potentially appear in a document with an associated probability distribution. Each company also has a probability distribution for the topics. The collection of topics is common across the universe of companies, but the probability distribution over topics different for each particular company. We can then imagine management writing its 10-k as a machine that randomly selects a topic based on the business' topic distribution, then randomly selects a word from that topic based on the probability distribution in the topic thus selected and repeats the process until enough words have been written down.

Topic modelling techniques allow us to infer the topics and the probability distributions from the business descriptions themselves. In other words, we work out what the topics are from the collection of documents, rather than having some pre-existing idea of what the topics might be.

Morningstar Equity Comparables then measures the similarity between the different managements' probability distributions to calculate a business similarity score. We assume similar probability distributions mean that management of different firms believe their companies do the same thing. Once

we have generated the probability distribution over the collection of topics, we compare the probability distributions to determine the similarity of the businesses. Similar probability distributions result in similar businesses.

**Methodology**

We start from the assumption that the managements of two businesses that do roughly the same thing will describe their businesses in the same terms. We then build a formal model of how managements go about writing descriptions of their own businesses. The model discovers a probability distribution of topics a business might talk about, where every business has its own probability distribution. We then discover the similarity between businesses as the similarity between topic distributions.

### Data Preparation

We use the business description from annual reports to determine company similarity. Our system is English-language only, however many companies from non-English speaking nations produce annual reports in English–these are included if they exist. We collect the documents as follows.
First, we look at major exchanges worldwide and produce a list of all currently listed common-share-like securities on those exchanges. The present version of the framework restricts to exchanges in the United States, Canada, United Kingdom, Australia, South Africa, and New Zealand, but we anticipate incorporating more exchanges over time. For these purposes, an ordinary share, a unit in a unit trust, and stapled securities are included as common-share-like securities, but ADRs are excluded. We then exclude all shares that have market caps of less than $10 million and that are registered on the Pink Sheets or OTCBB. We use the Morningstar Company ID for the share to identify the business.

For each business we find the most recent English-language annual report within the past 15 months as at the model generation date; if there is no annual report, we find the most recent prospectus, product disclosure statement, or offering document. We use the Morningstar Document Library as our source of base documents. Morningstar Document Library is a product that compiles fund and company disclosures from many exchanges worldwide, including all major exchanges, and tracks approximately 75,000 companies. If there is no English-language document published satisfying these criteria within this timeframe, we ignore the firm. We expect that a firm will have an annual report in a PDF, HTML, or ordinary text file format; if no such file exists, we ignore the company. In the present iteration approximately 10,000 companies satisfy our criteria. It is at this point that we also pull in all English Wikipedia data and parse the files to extract all publicly listed company information. Analyst reports are also downloaded from the Morningstar publishing system in all available companies. Adding these two data sources, we can increase coverage to around 14,000 companies.

For each document we extract a sequence of the words, including punctuation, in the document. For PDF documents we extract the text and tables in the first 80 pages of a document, perform automated column recognition, and discard tabular data, to construct a sequence of text. In PDF documents we perform some small modifications to the text to normalize ligatures and remove diacritics because the character encoding can be inconsistent. For HTML documents–mostly SEC filings–we identify and discard HTML tags. For text documents we split on whitespace. If our extractor can find fewer than 200

words from the document we ignore the company–this occasionally occurs where firms submit crude scans of paper filings, however it is rare.

**Bag of Words Representation**

Each document and each sentence has a "bag of words" representation, as is commonly used in the information retrieval community, and we use this basic representation frequently in our system. A bag of words representation describes each document or sentence as a vector (sequence of numbers) with one entry for each word that might possibly occur, and where the number in the vector corresponds to the number of occurrences in the text itself.

This depends on building a dictionary of all words that occur in the collection of disclosure documents. The dictionary is just a collection of words. We have a stoplist of very common English and finance words and remove these words from our dictionary. When constructing a bag of words representation of a document, if a word does not appear in the document but does appear in a dictionary compiled from all documents, then the word is assigned a value of 0.

We do not use the words themselves for our dictionary. We employ lemmatization on these words, a process that maps words to their common roots, so for instance the word "businesses" is mapped to "business." We use the WordNet lemmatizer. This has the effect of reducing the size of the collection of words from over 150,000 to around 50,000. In the sentence "Our business deals with many businesses," we would give the root "business" a count of 2.

Suppose our dictionary has 50,000 distinct words. For instance, the sentence "We currently own the drug delivery technology" would be encoded as a 1xN vector with approximately 50,000 zero entries and a 1 at the entries corresponding to the four words "currently," "drug," "delivery," and "technology." The words "we," "own," and "the" are common words in the stoplist.

**Exhibit 1** Sentence Vector Example

| | |
|---|---|
| abandon | 0 |
| … | |
| currency | 0 |
| currently | 1 |
| … | |
| delisting | 0 |
| delivery | 1 |
| … | |
| drowsiness | 0 |
| drug | 1 |
| … | |
| technique | 0 |
| technology | 1 |
| … | |
| zurich | 0 |

Source: Morningstar Inc.

We can assemble a collection of vectors into a term-document matrix ("term" being a generalization of the notion of "word"). In a term-document matrix, the $(i, j)$th entry contains the number of occurrences of the word indexed by $i$ in a column corresponding to document $j$. Each document uniquely identifies a company, so perhaps this could be thought of as a word-company matrix.

By itself, the term-document matrix is informative; it's reasonably simple to determine the line of business from the term-document matrix. Considering the following example fragment, it's reasonably straightforward to work out what company does what just from the word counts. For instance, for AVB, we see the words "real" and "estate" occur roughly equally, and the word "property" occurs a lot–and indeed, this organization is a REIT heavily involved in real estate.

**Exhibit 2** Term Document Matrix Example

| | AVB:USA | IBM:USA | KO:USA | MORN:USA | WFC:USA |
|---|---|---|---|---|---|
| asset | 25 | 8 | 1 | 36 | 24 |
| bank | 0 | 1 | 0 | 4 | 98 |
| branch | 0 | 0 | 0 | 0 | 3 |
| brand | 4 | 5 | 33 | 0 | 0 |
| consumer | 1 | 1 | 17 | 0 | 11 |
| estate | 26 | 0 | 0 | 0 | 0 |
| fdic | 0 | 0 | 0 | 0 | 32 |
| federal | 5 | 2 | 5 | 0 | 32 |
| fund | 26 | 0 | 3 | 67 | 13 |
| intellectual | 0 | 11 | 0 | 0 | 0 |
| interest | 21 | 0 | 14 | 0 | 11 |
| loss | 2 | 2 | 1 | 1 | 3 |
| property | 20 | 12 | 1 | 2 | 0 |
| rate | 1 | 5 | 4 | 5 | 18 |
| real | 26 | 2 | 0 | 0 | 0 |
| technology | 3 | 48 | 5 | 4 | 0 |

Source: Morningstar Inc.

Note that this is a fragment that displays more common entries. In a real-world example we typically see less than 3% of entries nonzero.

But in the example above, we can see the word "property" means different things for different businesses. The REIT AvalonBay (AVB) uses the word to refer to real estate. IBM typically uses the word in the context of intellectual property rights. These are different real-world objects described by the same words. We need a way to associate words together into a similar concept and find the underlying concepts or themes the businesses discuss. We will return to this notion in the topic model section below.

**Topic Model**

With two vectors in hand describing companies, we could calculate a similarity. But we employ a topic model for dimensionality reduction at this point. We use Latent Dirichlet Allocation[3] for a topic model and give a précis of the model here–this presentation follows the cited paper closely.

3 http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

A language model in this context is a probability distribution whose range is a collection of suitably encoded documents. Estimating a language model amounts to discovering important features of that probability distribution. The approach is Bayesian: In other words, we estimate the important parameters of the language model from some collection of documents that we model as being a sample from the distribution, and therefore we need to assume a prior distribution of the probability distribution parameters.

The language model describes a collection of high-probability documents inside the complete space of potential documents. We want to find the features of documents that make them high-probability documents, and then measure the similarity on these features, because these features represent something company managements wanted to emphasize when describing their companies.

Suppose we have some document $d$, which here is a sequence of integers, each integer coding for a word. Suppose we also have some preset number of words $N$. We then can construct a document in a two-step process:

1. Choose some $\theta \sim Dir(\alpha)$, that is, given some $\alpha$ that parameterizes a Dirichlet distribution, draw some $\theta$ from the distribution.
2. For each of the $N$ words $w_n$:
Choose a topic $z_n \sim Multinomial(\theta)$.
Choose a word $w_n$ from $p(w_n|z_n, \beta)$, a multinomial probability distribution.

Here $\beta$ is a $k \times V$ matrix where $\beta_{ij} = p(w^j = 1|z^i = 1)$–that is, the probability of seeing word $j$ given that we have drawn topic $i$ from the multinomial distribution $\theta$.

Note that multiple topics are drawn within the same document, so a document will in general contain multiple topics.

It is straightforward to see from this that given some $\alpha$ and $\beta$ it is possible to write down an expression for the joint distribution of the particular $\theta$, the sequence of $N$ topic selections $z$, and the selection of $N$ word selections $w$ as

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta).$$

We can then integrate over the $\theta$ and sum over $z$ to get the marginal distribution of a single document:

$$p(w|\alpha, \beta) = \int p(\theta|\alpha)\left[\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(w_n|z_n, \beta)\right] d\theta.$$

We can then describe the probability of a document collection $D$ as follows:

$$p(D|\alpha,\beta) = \prod_{d\in D}\int p(\theta_d|\alpha)\left[\prod_{n=1}^{N_d}\sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\right] d\theta_d.$$

Suppose that given some $D$, we can find a mechanism to identify $\beta$ and $\theta$ for that collection of documents. That is, suppose we have some mechanism to find which words are more or less likely given a particular topic and to find the most likely distribution of topics for the particular documents we have. By setting the number of topics to be substantially smaller than the number of words, we can effectively accomplish dimensionality reduction. Suppose we build a cross-tabulation of the topics estimated by the model and documents; for long documents, we expect that two documents with identical topic distributions $\theta$ will have near-identical rows in this cross-tabulation table. All the information about the company's activities appears in this topic distribution. So, if the topics are similar, companies are using similar word groups to describe their companies. There are many fewer topics than words, and the topics generalize away the specific activities of companies into the activities shared across companies The model above is very close to what we implement. In practice we fix $\alpha$ so that a typical firm will have material weights on three to five topics, which corresponds to an $\alpha$ of 0.1/k. We also employ a smoothing technique for $\beta$ estimation described in the original Latent Dirichlet Allocation paper, which effectively puts a prior on $\beta$; this is an implementation detail that does not affect the discussion here.

We want particular values of $\beta$ and $\theta$ that maximize the probability of seeing $D$ the sample of documents companies actually use to describe themselves. In practice we use a variation of the technique called Gibbs Sampling. Note that the word order is never used in this formalism, so two documents with only permuted word orders will have exactly the same probability–in practice we base our code on a bag of words representation of the data and calculate the probability of one particular instance of the bag of words. The details are outside the scope of this paper, but see the original paper on Latent Dirichlet Allocation for details. Our implementation is derived directly from this work with only minor changes.

**Measuring Company Similarity**

Having determined the topics that a business subscribes to, we need to determine how similar they are. We experimented with several divergence metrics and discovered in informal testing that KL divergence between the topic distributions provides a good divergence measure. If $\boldsymbol{x}$ and $\boldsymbol{y}$ are vectors of topic probabilities, then KL divergence is given by

$$KL(\boldsymbol{x},\boldsymbol{y}) = \sum_{1\le i\le k} x_i log\left(\frac{x_i}{y_i}\right).$$

We don't deal with the $y_i = 0$ case because all of our topics have a small nonzero weight. To make the measure symmetric, we use

$$\Gamma(\boldsymbol{x},\boldsymbol{y}) = \frac{KL(\boldsymbol{x},\boldsymbol{y}) + KL(\boldsymbol{y},\boldsymbol{x})}{2}$$

as our divergence measure between two firms. It is clear that $\Gamma(\boldsymbol{x},\boldsymbol{x}) = 0$. Typically, when $x_i$ and $y_i$ differ substantially either $x_i$ or $y_i$ will be close to 0. If it is $y_i$ that is small and $x_i$ that is close to 1, then

the $x\_i \log(x\_i/y\_i)$ term will be large. The corresponding term from the $KL(\boldsymbol{y}, \boldsymbol{x})$ expression will be close to 0. When both companies do not use a topic, both $x_i$ and $y_i$will be small and approximately equal, and so the sum of the $x\_i \log(x\_i/y\_i)$ and $y\_i \log(y\_i/x\_i)$ expressions are close to 0. If $x_i = y_i$ –that is, two businesses use a topic equally–then the sum of the expressions is 0. So the score means that if two firms both use a topic not much, or use a topic equally, then the topic causes little divergence. But if one company uses a topic but another does not, this will cause divergence between the firms. This gives the intuition for how one company discussing a topic and another not discussing the topic results in $\Gamma$ being large around that particular topic pair.
This $\Gamma$ is what we publish as our firm divergence.

### Local Maxima

We discovered through informal testing that topic models would often converge imperfectly, converging to local maxima. To avoid this problem, we build several topic models. Furthermore, we discovered empirically that topic models with low numbers of topics worked well to describe relationships between companies poorly correlated on market, whereas models with a large number of topics worked well to capture the correlation relationships for more loosely comparable companies. For this reason, we built models with a various number of topics. Through a series of experiments, we determined that three sets of models, with 150, 50, and 15 topics, for nine models altogether, worked well. To calculate divergence on the mixture model, we take the mean of the divergences calculated by all the models.

## Conclusion

We have shown a new, text- based system for determining the similarity between two companies. This system first collects diverse sets of text, then preprocesses and extracts clean data, builds a topic model to reduce the dimensionality of that data, and finally compares the company topics to come up with company-company scores. Multiple models are joined together to come up with more-robust scores. This bottom-up approach differs from the hierarchical systems used today and provides two main benefits: the ability to rank similarities and the ability to provide cross-sector scores.

**About Morningstar® Quantitative Research**

Morningstar Quantitative Research is dedicated to developing innovative statistical models and data points, including the Morningstar Quantitative Rating, the Quantitative Equity Ratings and the Morningstar Risk Model.

22 West Washington Street
Chicago, IL 60602 USA